

遅延時間を考慮したコンピューティング資源とネットワーク資源の同時割当て法

Proposed optimal simultaneous allocation of computing resources and communication resources

津村 重宏[†] 栗林 伸一[†] 田邊 正雄[‡]

Shigehiro TSUMURA[†], Shin-ichi KURIBAYASHI[†] and Masao TANABE[‡]

1. まえがき

筆者らは、ブロードバンド・ユビキタスネットワークサービスやグリッドコンピューティングサービスを前提に、各サービス要求に対してコンピューティング資源とネットワーク資源の両方を同時に効率よく割り当てる方式を提案した⁽¹⁾。

本稿では、さらに「遅延時間」という資源属性を追加し、その条件を満足することを前提に、両方の資源を同時に割り当てられない確率(NG率)を最小にする資源割当て法を提案する。

2. コンピュータ資源とネットワーク資源の同時割当て法

2.1 基本的な考え方

ブロードバンド・ユビキタスネットワークやグリッドコンピューティングサービスでは、様々なロケーションに分散するサーバ群(コンピューティング資源)、ネットワーク帯域(ネットワーク資源)、ストレージ資源、ファイアウォールなどのセキュリティ資源、など複数種類の資源の中から、全ての資源を同時に選択してサービスを提供する必要がある。本稿では文献(1)と同様に、まずコンピューティング資源とネットワーク資源の2つを取り上げ、両方の資源を同時に割当る前提で評価する。

通常、サービス要求毎に要求されるコンピューティング資源量とネットワーク資源量は均一でなく、サービス毎に異なる。しかし、データセンタなどに設置されたコンピューティング資源とそれに接続するネットワーク資源を別々に選択することはできない。このため、資源割り当てをうまく行わないと、片方の資源のみ空いたセンタが発生し、結果として両方の資源を必要とするサービスに対応できないというデッドロック状態(図1)が発生する。

また、サービス要求毎に、サービス開始から完了までの時間(サービス処理時間T)に制約がある。例えば、サービス処理時間を100ms以下としたいサービス要求に対しては、ネットワーク資源に空きがあっても、100ms以上の転送遅延がかかるものは選択できない。同様に、コンピューティング資源に空きがあっても、処理性能が低く100ms以上も計算時間がかかるものは選択できない。このように、仮に資源に空きがあっても、サービス条件を満足できないために、要求NGとなるケースを極力減らすことも考える必要がある。この点が、文献(1)の評価との違いである。

ネットワーク資源属性としては、遅延時間(例えば、地上回線と衛星回線であれば、200ms以上の差がある)、通信品質(ビット誤り率やパケット廃棄率などが大きいと再送が多くなり実行スループットも低下する)などが影響する。一方、コンピューティング資源属性では、計算処理時間(高速のCPUやマルチCPUを持つサーバであれば処理を短時間に実行できる)などが影響する。今回の評価では、ネットワーク資源の遅延時間だけを取り上げ、その制約を考慮した資源割り当てアルゴリズムを評価する。

資源の遅延時間だけを取り上げ、その制約を考慮した資源割り当てアルゴリズムを評価する。

2.2 前提条件

- 1) 例えばコンピューティング資源ではCPU100%を「100」、ネットワーク資源では150Mb/sを「100」として、両者の要求量の単位を合わせる。
- 2) 資源要求量は、要求量×保留時間(H)で求める。
- 3) ネットワーク資源では、遅延時間の違いを考慮する。
- 4) 同じサービス要求に対して、要求NG率が最も小さいアルゴリズムを最適とする。

2.3 資源同時割り当て法

文献(1)で提案した2案(案a,案b)に加え、案cを追加する。この案は、サービス処理時間Tに対する制約が厳しい要求のために、遅延時間の小さいネットワーク資源を極力空けておくものである(サービス処理時間の平均は長くなるが、要求NG率を小さくすることを優先)。

- ・案a. 複数ある資源を順番に選択(シーケンシャル選択)。
- ・案b. コンピューティング資源とネットワーク資源に対する要求資源量を比較し、資源要求量の大きい資源に着目。そして、その時点で最も空き量が少ない資源を選択(最繁群選択)。
- ・案c. 空き資源がありかつ許容サービス処理時間を満足するネットワーク資源の中で、最も遅延時間の大きい方を選択。

3. シミュレーション評価

3.1 評価条件

(1) 2つのセンタA, Bにコンピューティング資源とそこに接続するためのネットワーク資源(センタA向けとB向け)を考える。

(2) 2.3節で提案した選択アルゴリズムに従い、最初にどちらを選択するかを決める。最初に選択したセンタで要求に満足できない場合はもう一方のセンタを選択。両方のセンタともに要求を満足できない場合には要求NGとする。

(3) センタAおよびセンタB向けネットワーク資源の遅延時間をそれぞれTa, Tbとする。

3.2 シミュレーション評価

簡単なシミュレーション結果の例を表1に示す。この例では、各センタのコンピューティング資源総量ならびにネットワーク資源総量をそれぞれ20, Ta=0.1s, Tb=0.5s, 保留時間H=5sとした。

表1より、以下の諸点が明らかである。

- 1) パターン1のようなケースでは、案bは資源量だけ注目して遅延時間の短い方にまず詰め込むため、後で遅延時間に対して厳しい要求が来ても(資源量に空きはあるが遅延時間の制約から)NGになってしまう。なお、案aは両センタに要求が分散され、結果として資源に空きがあるにも関わらず遅延時間制約からNGとなる可能性が案bより低くなる。

[†]成蹊大学理工学部, Department of Computer and Information, Seikei University

[‡]日本電信電話株式会社 NTT情報流通プラットフォーム研究所, NTT Information Sharing Platform Labs, NTT Corporation

2)パターン2のようなケースでは、案b、案cともにNG率が低くなっている。これは、最初に遅延時間に対して厳しい要求が連続するため両案とも遅延時間制約の厳しいセンタに詰め込む結果になり、その後の要求は資源に空きのある遅延時間制約の厳しくないセンタを選択できるためである。案aは両センタを順番に選択するため、両センタ資源に虫食い状態が発生するため、NG率が高くなる。

3)パターン3のようなケースでは、案cはNGが発生しないが、案bでは遅延時間制約を考慮しないで詰め込むため、その後発生する遅延時間制約の厳しい要求に答えられないため、NGが発生しやすい。案aはパターン2と同様に、両センタ資源に虫食い状態が発生するため、NG率が高くなる。

4)パターン4のようなケースでは、案cは案bに比べて資源量の制約でNGとなる可能性が高くなり、不利となる。

以上から、案bは遅延時間条件は考慮せず使用率の高い資源に詰め込むように働くため、最初に遅延時間の短い資源を最初に選択してしまうと、後で発生するサービス処理時間制限の厳しい要求に答えられなくなり、案cに比べ不利となることがわかる。一方、

要求NGの理由がサービス処理時間でなく資源量が多いケースでは逆に案cが案bに比べ、不利となることがわかる。

4. むすび

今後、要求資源量、発生パターン、保留時間、サービス処理時間、などをパラメータとして、3案の適用領域の明確化を行う予定である。さらに、ストレージ資源など資源種別を3以上にした場合、センタ数を3以上にした評価も合わせて行う予定である。

<参考文献>

- (1)栗林他：“コンピューティング資源とネットワーク資源の同時最適割当て法”，電子情報通信学会全国大会2005.3
- (2)小川，栗林他：“B-ISDNにおける信号転送用VC設定法ならびに帯域選択法の提案”，電子情報通信学会 信学技報SSE94-182(1995-01)。

表1. シミュレーション結果

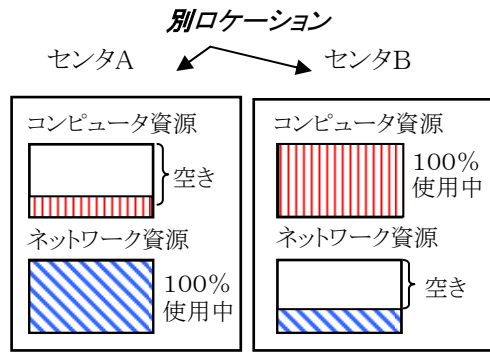


図1. 2つの資源が空いているにもかかわらず同時に選択できないケース(デッドロック状態)

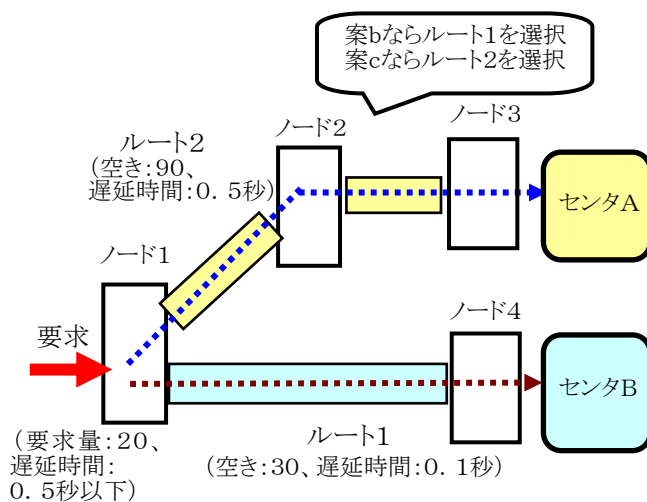


図2. 遅延時間という制約がある場合の資源選択例

【シミュレーション条件】

(1) センタA:コンピュータ資源(C)の総和20, ネットワーク資源の総和20

(2) センタB:コンピュータ資源(C)の総和20, ネットワーク資源の総和20

(3) 要求の発生

各要求は1秒間隔(一定)で発生. また、各要求の保留時間は5秒(一定)。

<パターン1>

「Cを6要求, Nを6要求, 許容遅延時間0.1秒」,

「Cを6要求, Nを6要求, 許容遅延時間0.5秒」,

「Cを6要求, Nを6要求, 許容遅延時間0.5秒」の繰り返しパターンで発生。

各要求は1秒間隔(一定)で発生。

<パターン2>

「Cを3要求, Nを3要求, 許容遅延時間0.1秒」,

「Cを3要求, Nを3要求, 許容遅延時間0.1秒」,

「Cを10要求, Nを10要求, 許容遅延時間0.5秒」の繰り返しパターンで発生。

<パターン3>

「Cを10要求, Nを10要求, 許容遅延時間0.1秒」,

「Cを5要求, Nを5要求, 許容遅延時間0.5秒」,

「Cを5要求, Nを5要求, 許容遅延時間0.5秒」の繰り返しパターンで発生。

<パターン4>

「Cを5要求, Nを5要求, 許容遅延時間0.1秒」,

「Cを5要求, Nを5要求, 許容遅延時間0.1秒」,

「Cを5要求, Nを5要求, 許容遅延時間0.5秒」,

「Cを5要求, Nを5要求, 許容遅延時間0.5秒」の繰り返しパターンで発生。

【結果】

シミュレーション結果を以下に示す. 要求発生パターン毎に各案の要求NG率を%で示す。

	案a	案b	案c
<パターン1>	0%	14%	0%
<パターン2>	17%	3%	0%
<パターン3>	17%	17%	0%
<パターン4>	20%	0%	20%